Санкт-Петербургский государственный университет

Математическое обеспечение и администрирование информационных систем
Информационные системы и базы данных

Дубатовка Алина Дмитриевна

Автоматическая генерация тонально аннотированных русскоязычных словарей для произвольной предметной области

Магистерская диссертация

Научный руководитель:
к. ф.-м. н., доцент Михайлова Е. Г.

Рецензент:
к. т. н., старший научный сотрудник Браславский П. И.

Санкт-Петербург
2017

SAINT-PETERSBURG STATE UNIVERSITY

Software and Administration of Information Systems
Information Systems and Databases

Alina Dubatovka

Automatic generation of domain-specific sentiment Russian dictionaries

Master's Thesis

Scientific supervisor:
Ph. D., associate professor Elena Mikhailova

Reviewer:
Ph. D., senior researcher Pavel Braslavski

Saint-Petersburg
2017

Оглавление

Введение

В современном мире задача анализа мнений вызывает интерес далеко не только у исследователей текстов на естественном языке. Автоматический анализ тональности текста находит своё применение в политике, бизнесе, киноиндустрии и актуален не только для политиков или компаний, но и для простых пользователей. В связи с большим разнообразием сфер применения актуальной становится не просто задача анализа мнений, но анализа отзыва конкретной предметной области. Большая часть алгоритмов для анализа мнений основана или хотя бы использует внутри себя словари положительных и отрицательных слов, поэтому их автоматическая генерация как для различных языков, так и для предметных областей является важной задачей, решаемой исследователями во всём мире.

В данной работе описывается алгоритм создания тонально аннотированного русскоязычного словаря для заданной предметной области, основанный на построении и анализе графовой модели. Было проведено исследование зависимости качества словаря от различных параметров построения и анализа графа. Важно отметить, что описываемый алгоритм не требует никакой предварительной разметки, а только достаточно большой корпус русскоязычных текстов из рассматриваемой предметной области, который несложно подготовить самостоятельно. Наш алгоритм не имеет жёсткой привязки к русскому языку и при желании может быть обобщён для создания словарей на других языках.

Создание словарей положительных и отрицательных слов происходит при помощи анализа графа, построенного на основе неразмеченного корпуса русскоязычных текстов, взятых из предметной области. Подобные корпусы могут быть сформированы автоматически для большинства областей. Вершинами графа являются прилагательные, а рёбра соединяют те из них, которые хотя бы в одном предложении были соединены между собой одним из сочинительных союзов.

Для разделения множества вершин этого графа на подмножества положительных и отрицательных слов выбираются небольшие началь-

ные множества, состоящие из прилагательных, эмоциональная окраска которых не вызывает вопросов вне зависимости от контекста и предметной области (например, “плохой”, “хороший”, “ужасный”, “превосходный”). Далее слова распределяются итеративно: каждый раз в множество положительных или отрицательных слов добавляется вершина, наиболее сильно связанная с уже имеющимися в нём вершинами. Было проведено сравнение нескольких весовых функций на рёбрах, а также функций расстояния для составления наиболее точных словарей. Данная работа также исследует применимость подхода, описанного в [8], для анализа тональности прилагательных в русскоязычных отзывах.

1. Обзор предметной области

Несмотря на активный интерес к рассматриваемой теме и большое количество работ в области анализа мнений, составление тонально аннотированных словарей не теряет своей актуальности. Основных причин для такой популярности две: необходимость подобных словарей для различных языков и зависимость эмоциональной окраски слова от рассматриваемой предметной области. Таким образом, задача построения словарей эмоциональной окраски для конкретного языка и предметной области является весьма актуальной, поскольку даже при наличии большого количества общедоступных размеченных данных, большая часть из них составлена для английского языка и рассматривает либо отзывы на фильмы, либо отзывы на технику.

Например, в [1] описан алгоритм составления одного из двух известных нам публично доступных русскоязычных словарей оценочных слов для мета-области товаров с помощью обучения ряда классификаторов на одной предметной области и последующего переноса полученной модели на другие предметные области. Далее, в [10] авторы пытаются уточнить полученный словарь с помощью анализа соответствующего подграфа RuThes тезауруса.

Задача составления тонально аннотированных словарей актуальна не только для русского, но и для многих других языков. Так, в [3] предлагается метод для автоматического составления словаря для новых языков (немецкий, русский, итальянский, французский, арабский и чешский) методом триангуляции с использованием составленных вручную словарей на английском и испанском языках. А в [2] исходный, составленный вручную словарь для немецкого языка расширялся за счёт построения графа на основе неразмеченного немецкого корпуса, как описано в [8], и его дальнейшего анализа с помощью классификации методом максимальной энтропии.

Различные графовые модели широко используются для таких подзадач как адаптация модели к новой предметной области, выделение тонально окрашенных предложений из текста или ранжирование слов

по степени эмоциональности. Авторы [13], имея корпус размеченных документов одной предметной области и неразмеченный корпус из другой предметной области, определяют полярность, строя и анализируя взвешенный граф, построенный на отзывах, где весом ребра является косинусная мера схожести документов. А в [11] рассматривается задача автоматического выделения тонально окрашенных предложений из текста с помощью поиска минимального разреза в графе, построенном на предложениях, содержащихся в одном документе, и связях между ними. Поскольку в таких областях, как анализ социальных сетей или сети интернет графы играют ключевую роль, разработаны различные алгоритмы их анализа, которые могут быть применены и к рассматриваемой в данной работе предметной области. Например, в [5, 6] предпринимается попытка адаптировать различные алгоритмы случайных блужданий (в частности, PageRank) к графу, построенному на основе eXtended WordNet [7], для ранжирования эмоциональной окраски слов.

2. Мотивация

Несмотря на то, что анализ мнений является активно развивающейся областью с широким кругом применения от отзывов на товары до социальных исследований и анализа блогов, а словарные подходы играли и продолжают играть большую роль в этой области (например, в таким системах как SentiStrength), для русского языка на данный момент известно всего два публично доступных словаря оценочной лексики. Причём первый из них, ProdSentiRus, составленный И. Четвёркиным и Л. Лукашевич [1], представляет собой список оценочных слов без указания знака их тональности (позитивной или негативной).

Второй словарь [9], составленный интернет-лабораторией LINIS, является полноценным тонально-аннотированным словарём для анализа социально-политических текстов на русском языке. Помимо словаря оценочной лексики данный ресурс содержит также корпус размеченных текстов социально-политической направленности, относительно которых и определялись эмоциональные оценки слов, составляющих словарь. Таким образом ресурс предоставляет возможность тестировать алгоритмы генерации словарей методом сравнения с "золотым стандартом", а также качество извлекаемых словарей для сентимент анализа имеющихся текстов.

Было проведено сравнение полученных в ходе данной работы словарей с описанными выше, а также анализ соответствия последних нашим данным. Для оценки согласованности словарей был вычислен коэффициент корреляции Пирсона. Причём в случае со словарём ProductSentiRus исследовалась согласованность отделения нейтральных слов от эмоционально окрашенных, поскольку данный словарь представляет только оценку эмоциональности слов, но не определяет знак тональности (положительный или отрицательный). Для этого словари LINIS и Hotels были приведены к бинарной шкале: как положительные, так и отрицательные слова получали оценку 1, в то время как нейтральные слова были отмечены оценкой 0. В случае сравнения словаря Hotels, полученного в ходе данной работы, со словарём LINIS вычислялась согласован-

ность разделения прилагательных на положительные, отрицательные и нейтральные. При этом для словаря LINIS, имеющего изначально пятибалльную шкалу от -2 до 2, не делалось различия между положительными (имеющими оценку 1) и очень положительными (имеющими оценку 2) словами — все они при сравнении получали оценку 1. Аналогично все негативные слова получали оценку -1. Полученные результаты представлены в таблице 1.

Таблица 1: Согласованность между различными тонально аннотированными русскоязычными словарями

	LINIS	ProductSentiRus	Hotels
LINIS	–	0.05	0.74
ProductSentiRus	0.05	–	0.08
Hotels	0.74	0.08	–

Кроме того, в таблице 2 приведены данные о покрытии каждого из словарей словами другого словаря, то есть отношение количества слов, содержащихся в обоих словарях к размеру словаря. Стоит отметить, что в данной работе рассматриваются алгоритмы, работающие только с прилагательными, поэтому все полученные в ходе экспериментов словари состоят только из прилагательных. Поэтому при вычислении покрытия из словарей ProductSentiRus и LINIS были предварительно удалилены все слова, представляющие другие части речи.

Таблица 2: Покрытие тонально аннотированных словарей

	LINIS	ProductSentiRus	Hotels
LINIS	–	0.36	0.34
ProductSentiRus	0.24	–	0.3
Hotels	0.65	0.49	–

Для оценки соответствия словарей данным, используемым в данной работе, из рассматриваемых текстов были выделилены все прилагательные и посчитано, какая часть из них покрывается каждым из упомянутых выше словарей. То же самое было проделано с текстами, предоставленными LINIS, для оценки того, насколько вообще словари

одной предметной области могут быть перенесены на другую. Таблица 3 содержит описанные результаты.

Таблица 3: Покрытие тонально аннотированными словарями рассматриваемых текстов

Источник текстов	LINIS	ProductSentiRus	Hotels
LINIS blogs	0.3	0.2	0.34
Hotels reviews	0.22	0.16	0.4

Приведённые выше результаты наглядно показывают, что даже наличие тонально аннотированных словарей, составленных для одной предметной области (которых для русского языка, на самом деле, практически нет), не решает проблемы анализа тональности текста, взятого из другой предметной области, поскольку подобные тексты отличаются не только набором используемых слов, но также эмоциональной окраской общеупотребительных лексических единиц. В связи с этим встаёт необходимость разработки алгоритмов автоматической генерации тонально аннотированных словарей для каждой конкретной предметной области или конкретного корпуса текстов, решению которой и посвящена данная работа.

3. Описание метода

Цель данной работы заключается в разработке такого метода построения тонально аннотированного словаря, который работал бы одинаково хорошо вне зависимости предметной области. Поэтому данный алгоритм не должен опираться ни на какие априорные знания о предметной области, а использовать только информацию, полученную из имеющихся данных — коллекцию неразмеченных текстов определённой тематики. Для достижения этой цели из данных извлекаются синтаксические связи между прилагательными, встречающимися в текстах, и на их основе строится графовая модель для последующего анализа.

Как показано в [8], сочинительные союзы, соединяющие однородные прилагательные и наречия, передают соотношение полярности соединяемых слов. Например, соединительные союзы, как правило, ставятся между словами, имеющими одинаковую полярность (“Вкусный и полезный завтрак”), а противительные — между словами с противоположной тональной окраской (“Дешёвый, но хороший отель”).

Данные отношения между эмоциональной окраской слов позволяют построить взвешенный граф, вершинами которого являются прилагательные, а рёбрами — связи между ними, помеченные количеством случаев, когда слова были соединены соединительным союзом, а когда — противительным.

Анализ конфигурации полученного графа позволяет оценить “положительность” или “отрицательность” слов, являющихся его вершинами, — чем лучше связана вершина с другими “положительными” вершинами и хуже с “отрицательными”, тем “положительнее” слово, ей соответствующее.

Таким образом, алгоритм построения тонально аннотированного словаря на основе исходного неразмеченного корпуса состоит из двух основных этапов, описанных далее: построение графа отношений между словами и его обработка.

3.1. Построение графа

Как было описано выше, вершинами в строящемся графе являются прилагательные, а рёбрами – сочинительные и противительные связи между ними. Чтобы построить такие рёбра, из текстов извлекаются однородные прилагательные и рассматривается, как они связаны между собой. Однородными считались прилагательные, согласованные в роде, числе и падеже, и удовлетворяющие шаблону

$$(ADV \mid NEG) * ADJ(, ?(AND \mid BUT)?(ADV \mid NEG) * ADJ)+,$$

где AND – соединительный союз “и”, BUT — один из противительных союзов (“но”, “да”, “зато”, “однако”), NEG — отрицательная частица “не”, ADV — наречие меры и степени (“очень”, “совсем”, “слишком”, “вполне”), а ADJ — собственно, прилагательное.

Для каждой пары из выделенного ряда однородных прилагательных формируется тональная связь между ними (положительная или отрицательная в зависимости от союза), которая впоследствии будет учтена при вычислении веса ребра. Например, из фразы “Вкусный, обильный, но не очень разнообразный и дорогой завтрак” получаются три положительные связи (вкусный, обильный), (вкусный, разнообразный), (обильный, разнообразный), а также три отрицательные — (вкусный, дорогой), (обильный, дорогой), (разнообразный, дорогой).

Для определения части речи, а также лемматизации слов используется морфологический анализатор русского языка Mystem [12]. Он работает на основе словаря и приводит слова к начальной форме (мужской род, единственное число, именительный падеж), а также печатает для них грамматическую информацию. Для слов, не входящих в словарь, строятся гипотетические разборы слов.

В случае, когда перед прилагательным встречается отрицательная частица “не”, знак связи слова меняется на противоположный. Например, в предложении “Бассейн большой, но не очень глубокий”, прилагательные “большой” и “глубокий” получат положительную связь, означающую совпадение тональной окраски, несмотря на то что соединены

противительным союзом “но”. Аналогично, при анализе фразы “Вкусная и совсем не дорогая кухня” слова “вкусный” и “дорогой” получат отрицательную связь и, следовательно, противоположную полярность.

Поскольку в качестве данных для анализа рассматриваются не литературные тексты, а отзывы пользователей сети Интернет, необходимо учитывать не только пунктуационные правила русского языка, но и наиболее часто встречающиеся (пусть и ошибочные) формы. Так, например, далеко не всегда пользователи ставят запятую, даже если правила предполагают её использование: разделение бессоюзных однородных членов, запятая перед “но” или между повторяющимися союзами “и”, поэтому шаблон не должен быть очень строгим.

3.1.1. Частица и приставка “не”

Помимо того что отрицание может быть выражено отдельно стоящей частицей “не”, оно также может появиться в виде префикса “не-” у прилагательного. Например, “не очень красивый”, “некрасивый” и даже “совсем не красивый”, по большому счёту, представляют собой одно и то же отрицание прилагательного “красивый”. Поэтому решено было проверить, как отрицательная приставка “не-” влияет на эмоциональную окраску прилагательного. Для этого были реализованы и сравнены два подхода: в одном прилагательные, скажем, “некрасивый” и “красивый” считались двумя разными вершинами графа, в другом же приставка “не-”, если это было возможно (то есть слово без приставки опознавалось Mystem’ом) отделялась и рассматривалась как отрицательная частица “не”, то есть прилагательное “некрасивый” приравнивалось к фразе “не красивый” и отдельной вершины для слова “некрасивый” не создавалось. На рисунках 1 и 2 представлены фрагменты графов, полученных без дополнительной обработки отрицаний и с удалением приставок “не-” и частиц “не” соответственно.

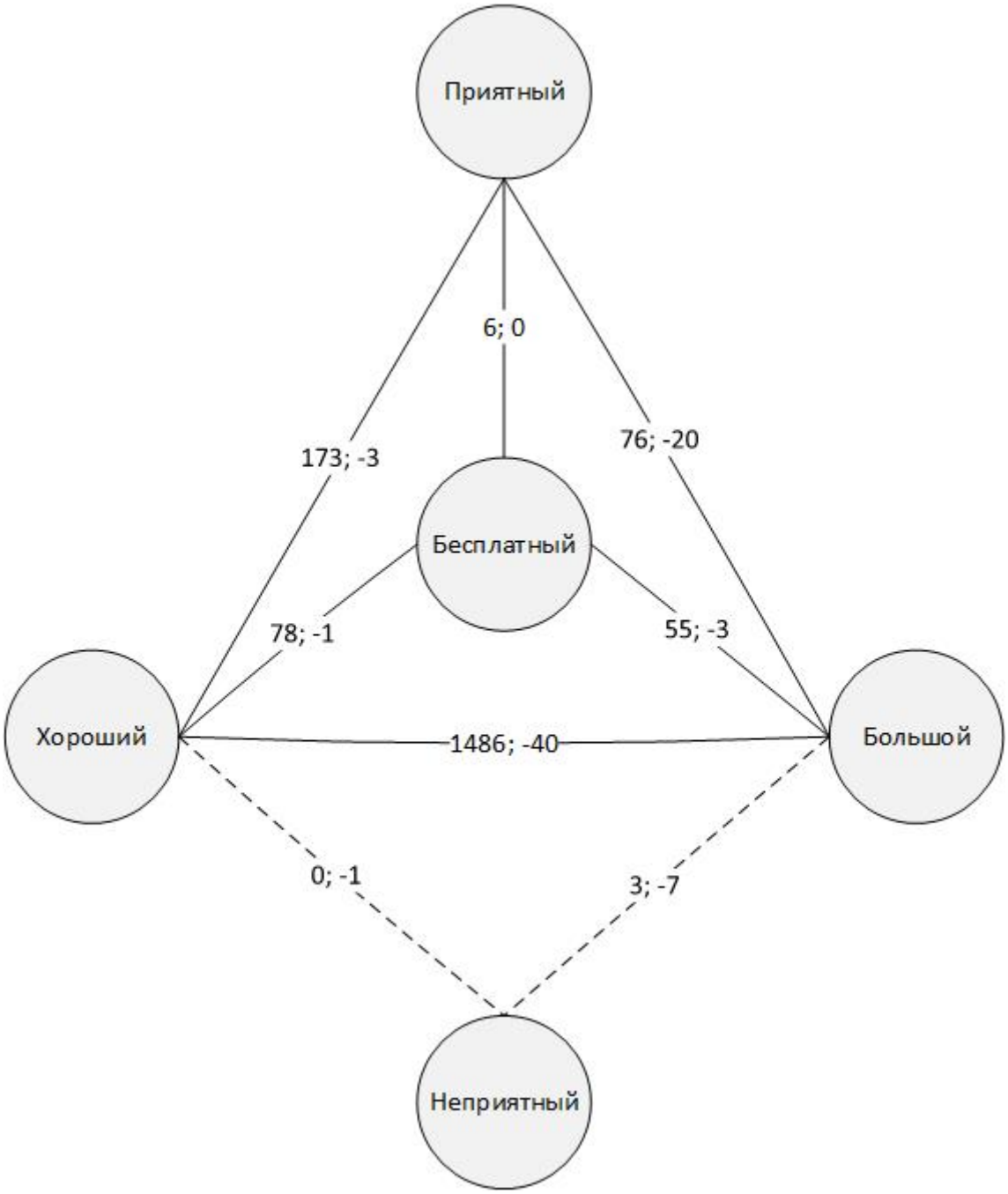

Рис. 1: Фрагмент графа, полученного без удаления “не-”

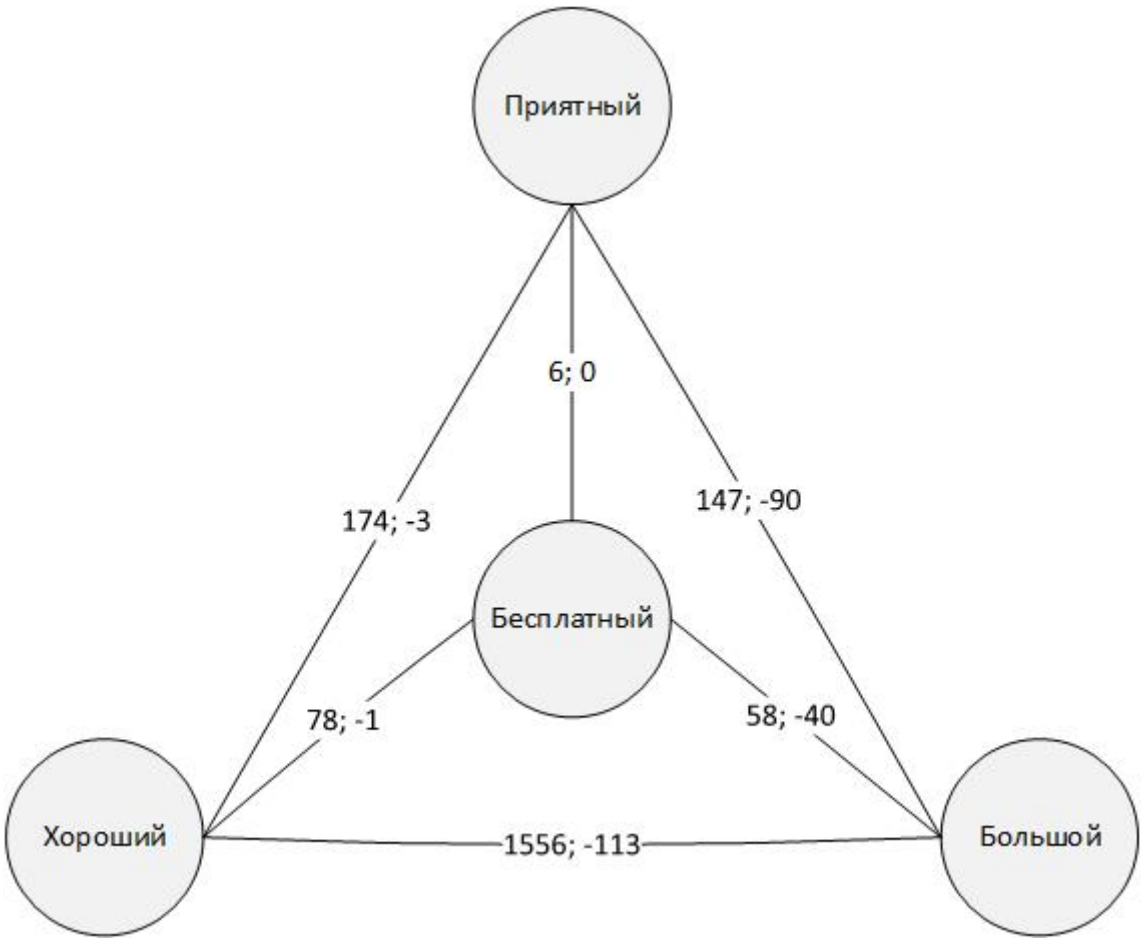

Рис. 2: Фрагмент графа, полученного с помощью удаления “не-”

3.2. Обработка графа

Полученный на предыдущем этапе граф необходимо разделить на два кластера: множество позитивных и негативных слов. Для этого в каждое из этих множеств по очереди добавляется, если это возможно, по одному кандидату из ещё не распределённых вершин. Кандидат выбирается по принципу наименьшего расстояния до рассматриваемого на данном шаге множества (“позитивного” или “негативного”). Множество кандидатов состоит из вершин, имеющих хотя бы одно ребро, ведущее в одно из этих итоговых множеств. После добавления очередной вершины в множество все её ещё не распределённые соседи добавляются в множество кандидатов, а расстояния от вершин-кандидатов до итоговых множеств пересчитываются.

3.2.1. Инициализация

Инициализировать позитивные и негативные множества позволяет тот факт, что эмоциональная окраска некоторых слов очевидна вне зависимости от контекста и предметной области. Поэтому изначально множество “положительных” слов состояло из таких очевидно позитивных прилагательных, как “хороший”, “отличный”, “замечательный”, “прекрасный”, “лучший”, а множество “отрицательных” слов — из негативных прилагательных “плохой”, “ужасный”, “отвратительный”, “отвратный”, “худший”, “убогий”.

3.2.2. Веса рёбер

Поскольку в ходе построения графа каждое ребро помечается парой чисел — количеством положительных и отрицательных связей соответственно, — встаёт вопрос о вычислении веса ребра на их основе. Это может быть как простая разность, так и произвольная линейная комбинация или вообще нелинейная функция. В описываемых экспериментах вес ребра вычислялся по формуле

$$weight(word_1, word_2) = \#(word_1\ AND\ word_2) - K \cdot \#(word_1\ BUT\ word_2)$$

, где K — коэффициент значимости противительных союзов. Данный коэффициент служит для придания большего веса противительным связям, которых в русскоязычных текстах гораздо меньше, чем соединительных, из-за разницы в частоте употребления соответствующих союзов.

3.2.3. Расстояния до итоговых множеств

Аналогичная проблема возникает и при определении функции расстояния от вершины до одного из итоговых множеств. Помимо того что каждую вершину с множеством может соединять несколько рёбер различного веса (что лучше: одно “тяжёлое” ребро или много “лёгких”?), она может ещё иметь рёбра, идущие в противоположное множество, которые также необходимо учитывать. Было проведено сравнение двух следующих наиболее распространённых и интуитивно понятных способов вычисления расстояния.

3.2.4. Самое тяжёлое ребро

Расстоянием до множества считается вес самого “тяжёлого” ребра, ведущего в это множество, то есть ребра с максимальным весом.

3.2.5. Сумма весов рёбер

Из суммы весов рёбер, ведущих в рассматриваемое множество, вычитается сумма весов рёбер, ведущих в противоположное множество — это и есть вес ребра.

3.2.6. Отделение нейтральных слов

Одной из основных проблем в задаче определения полярности является проблема отделения нейтральных слов от тонально окрашенных. Поскольку описанный алгоритм делит граф на две части, в результате формируются только списки позитивных и негативных прилагательных, которые могут, хотя и не должны, включать в себя и

нейтральные слова. Последние встречаются как правило чаще эмоционально окрашенных, а также могут появляться в тексте в паре как с положительными, так и с отрицательными прилагательными. Поэтому вершины, представляющие такие слова, часто имеют высокие степени и бывают сильно связаны с положительным и отрицательным кластерами. Из-за этого при разбиении графа такие прилагательные могут ошибочно попасть в списки эмоционально окрашенных слов. В нашем методе на это влияет момент остановки алгоритма, когда вершины перестают добавляться в множества, поскольку слова с ярко выраженной эмоциональной окраской обычно имеют прочную связь только с одним из кластеров и добавляются в него в первую очередь. В данной работе исследуется скорость деградации словаря при увеличении количества шагов алгоритма и, соответственно, размера словаря.

4. Описание экспериментов

4.1. Описание данных

На вход алгоритмам подавались обезличенные отзывы об отелях без предварительной разметки, из которых впоследствии выделялись соединительные и противительные связи прилагательных и строился граф для дальнейшего анализа с целью построения словарей позитивных и негативных слов. Таблица 4 содержит описание тестовой коллекции, а также характеристики графов, построенных с помощью и без удаления приставки “не-”.

Таблица 4: Характеристики исходного датасета и построенных графов

Характеристика	Без удаления “не-”	После удаления “не-”
Количество отзывов	259023	259023
Количество различных прилагательных	11716	11716
Количество связей	2879553	2879553
Количество рёбер	986349	939536
Количество вершин	10665	10114

4.2. Точность

Для того чтобы оценить точность работы алгоритмов на всех данных, результат был размечен на три категории: позитивные, негативные и нейтральные. Таким образом был получен “большой” словарь положительных, отрицательных и нейтральных слов, с помощью которого оценивали точность алгоритмов. Этот и есть тот самый словарь Hotels, который сравнивался со словарями LINIS и ProductSentiRus в начале данной работы. В данном случае оценивалась не только классическую точность, но и точность отделения положительных слов от отрицательных без учёта нейтральных слов, поскольку выделение нейтральных слов на деле является самостоятельной задачей [1, 10]. Таблица 5 содержит размеры “большого” тонально аннотированного словаря, а также

словарей, полученных в результате работы алгоритмов с учётом и без учёта приставки “не-”.

Таблица 5: Размеры “большого” и автоматически сгенерированных словарей

	Позитивный словарь	Негативный словарь	Нейтральный словарь	Всего
Без удаления “не-”	5252	2815	–	8067
С удалением “не-”	4936	2695	–	7631
“Большой” словарь	1948	1946	4951	8845

4.3. Оценка полноты

Поскольку из-за большого объёма входных данных, разметить все прилагательные вручную не представляется возможным, полнота была оценена следующим образом. Из всех имеющихся данных вручную были размечены 500 отзывов: все встречающиеся там прилагательные были разделены на позитивные и негативные в зависимости от того, с какой эмоциональной окраской они встречались в отзыве. Таким образом был составлен “ручной” тонально аннотированный словарь позитивных и негативных слов. Далее на основе этого словаря была посчитана классическая полнота — полученная величина принималась за оценку полноты построенных алгоритмом словарей. Поскольку 500 отзывов для оценки были выбраны случайно, а распределение прилагательных по отзывам предполагается равномерным, можно считать рассматриваемую выборку несмещённой. Таким образом полнота, посчитанная по “ручному” словарю, составленному на основе из этих 500 отзывов, приближает полноту на всех данных. Таблица 6 содержит размеры “ручного” словаря, а также так называемых “малых” словарей — результата пересечения “ручного” словаря со сгенерированными словарями.

Таблица 6: Размеры “ручного” и “малых” автоматических словарей

	Позитивный словарь	Негативный словарь	Всего
Без удаления “не-”	164	74	238
С удалением “не-”	163	83	246
“Ручной” словарь	173	127	300

4.4. Исследование скорости деградации словаря

Для исследования скорости деградации словаря и зависимости качества результата от момента остановки был построен график значения величины $Precision@n$, где $Precision@n$ — точность первых n слов из каждого словаря.

Также для исследования зависимости качества словаря от коэффициента значимости отрицательных связей K был построен график зависимости F_1-меры от значения K.

5. Результаты

Таблицы 7 и 8 содержат результаты работы алгоритмов без удаления приставки “не-” и с ним соответственно.

Таблица 7: Результаты работы алгоритма без удаления “не-”

Метрика	Позитивный словарь	Негативный словарь	Общий словарь
Полнота	0.94	0.64	0.82
Точность	0.33	0.63	0.42
Точность без учёта нейтральных	0.82	0.91	0.86
F_1-мера	0.48	0.75	0.57
F_1-мера без учёта нейтральных	0.88	0.75	0.84

Таблица 8: Результаты работы алгоритма после удаления “не-”

Метрика	Позитивный словарь	Негативный словарь	Общий словарь
Полнота	0.95	0.71	0.86
Точность	0.29	0.59	0.37
Точность без учёта нейтральных	0.76	0.90	0.82
F_1-мера	0.44	0.65	0.53
F_1-мера без учёта нейтральных	0.84	0.80	0.84

На рисунках 3–6 и 7–10 представлены графики величины $Precision@n$ для положительных и отрицательных словарей, составленных в результате обработки всех отзывов, соответственно. Несложно заметить, что словари начинают очень быстро деградировать за счёт включения нейтральных слов, однако, деградация “отфильтрованных” словарей, содержащих только окрашенные слова, происходит гораздо медленнее. Также стоит отметить, что добавление или удаление префикса “не-” почти не влияет на поведение графиков, в случае, когда нейтральные слова учитываются, однако, для отфильтрованных словарей даёт прирост

в точности, начиная с длины 500, хотя сначала приводит к падению точности.

На рисунках 11 и 12 представлены графики зависимости F_1-меры с учётом и без учёта нейтральных слов от параметра K для позитивных и негативных словарей. Рисунки 13 и 14 содержат точечную диаграмму полноты (ось абсцисс) и точности (ось ординат) при разных значениях K от 1 до 10 (точки подписаны соответствующими значениями параметра).

Рис. 3: Без удаления “не-” с учётом нейтральных

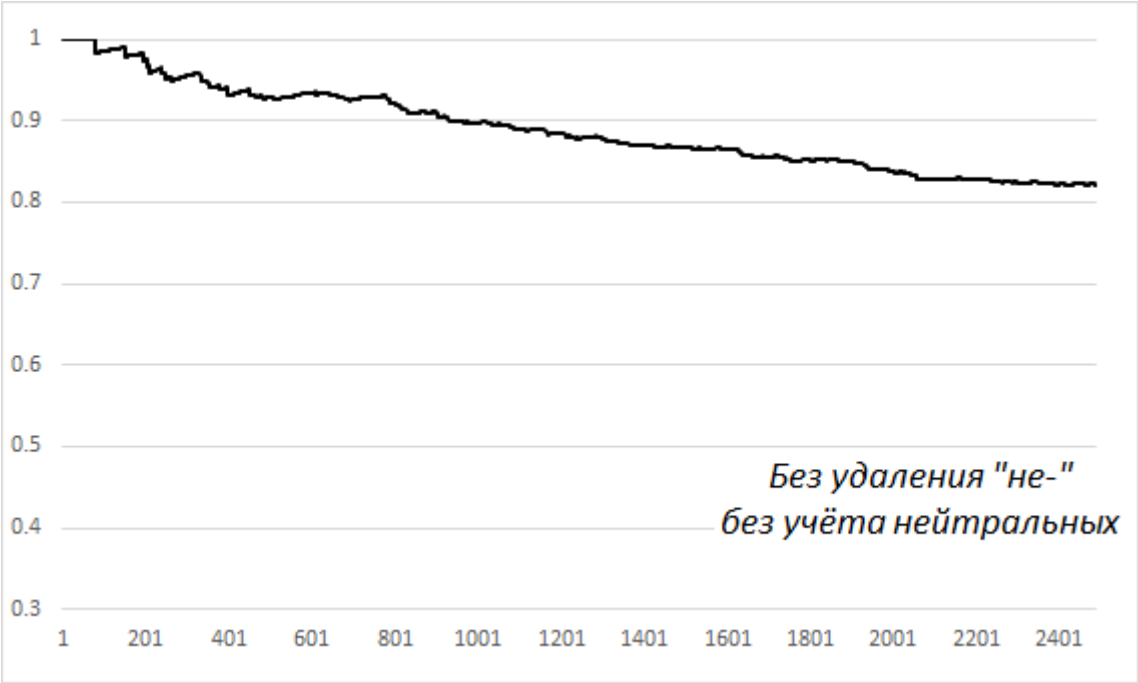

Рис. 4: Без удаления “не-” без учёта нейтральных

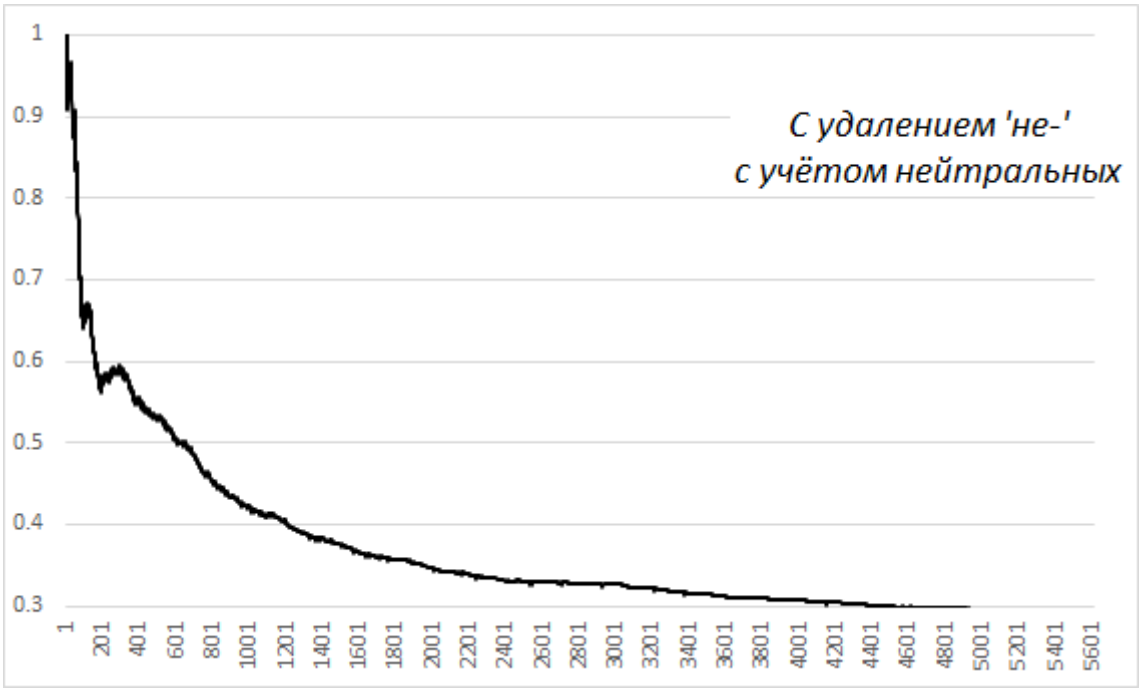

Рис. 5: С удалением “не-” с учётом нейтральных

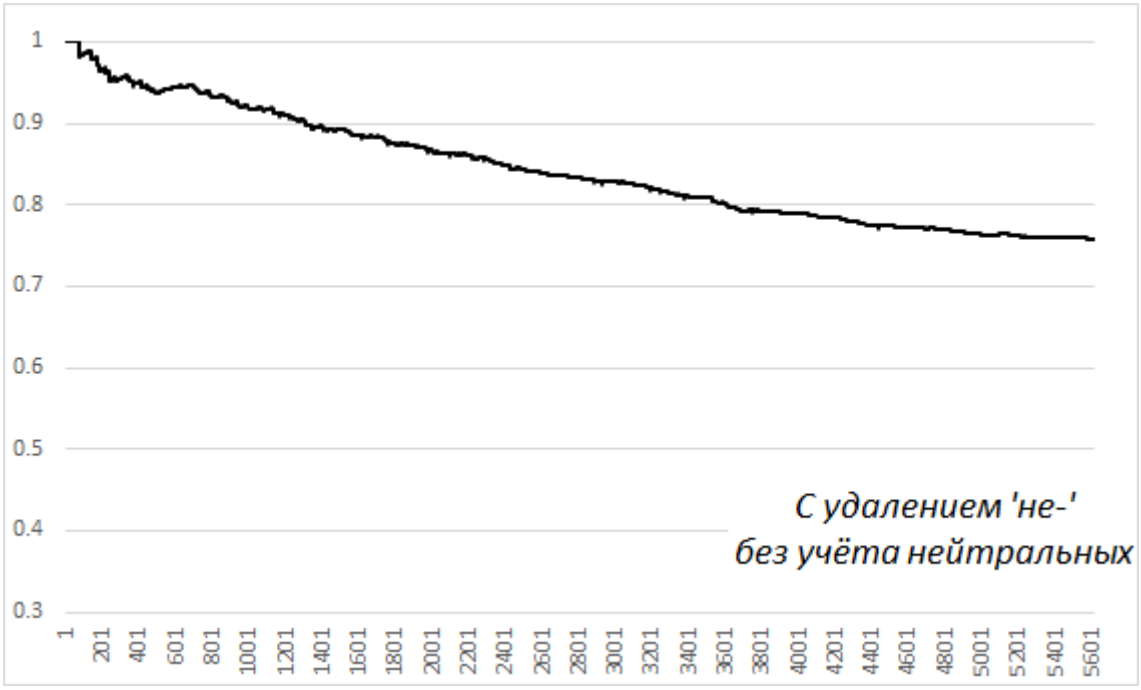

Рис. 6: С удалением “не-” без учёта нейтральных

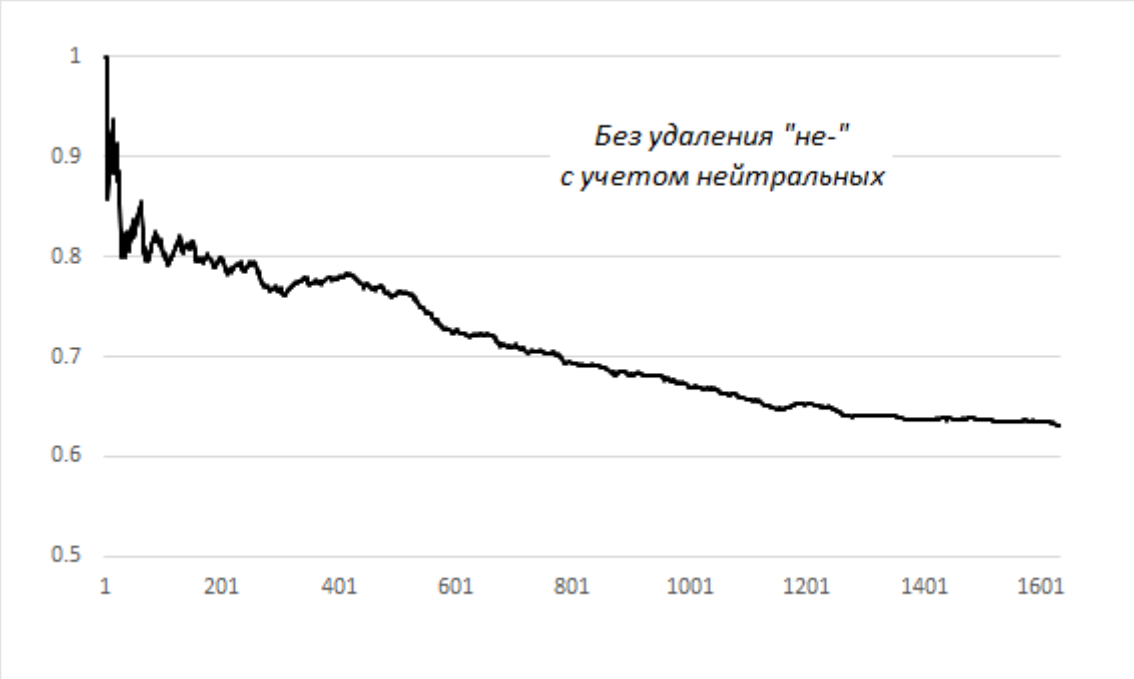

Рис. 7: Без удаления “не-” с учётом нейтральных

Рис. 8: Без удаления “не-” без учёта нейтральных

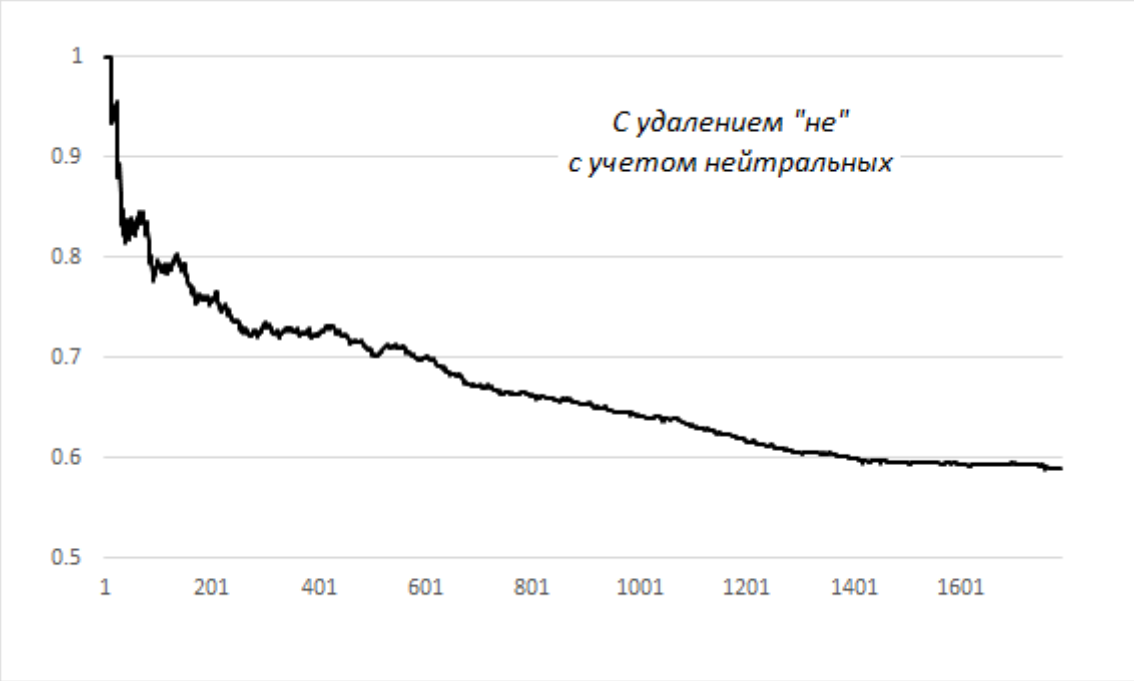

Рис. 9: С удалением “не-” с учётом нейтральных

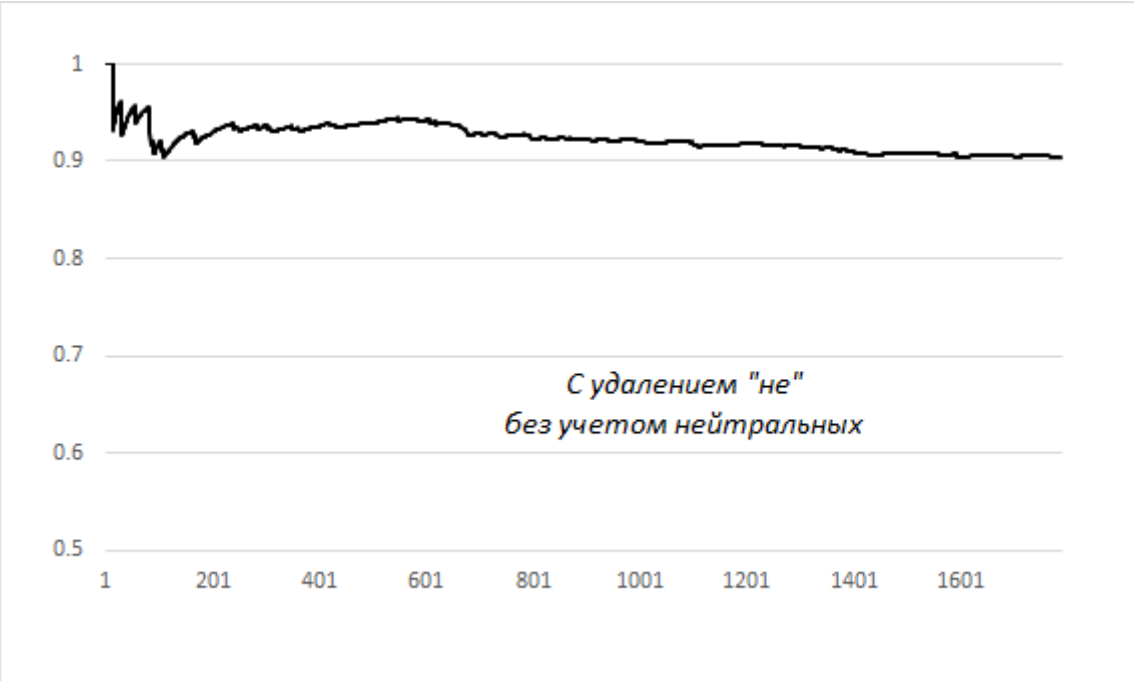

Рис. 10: С удалением “не-” без учёта нейтральных

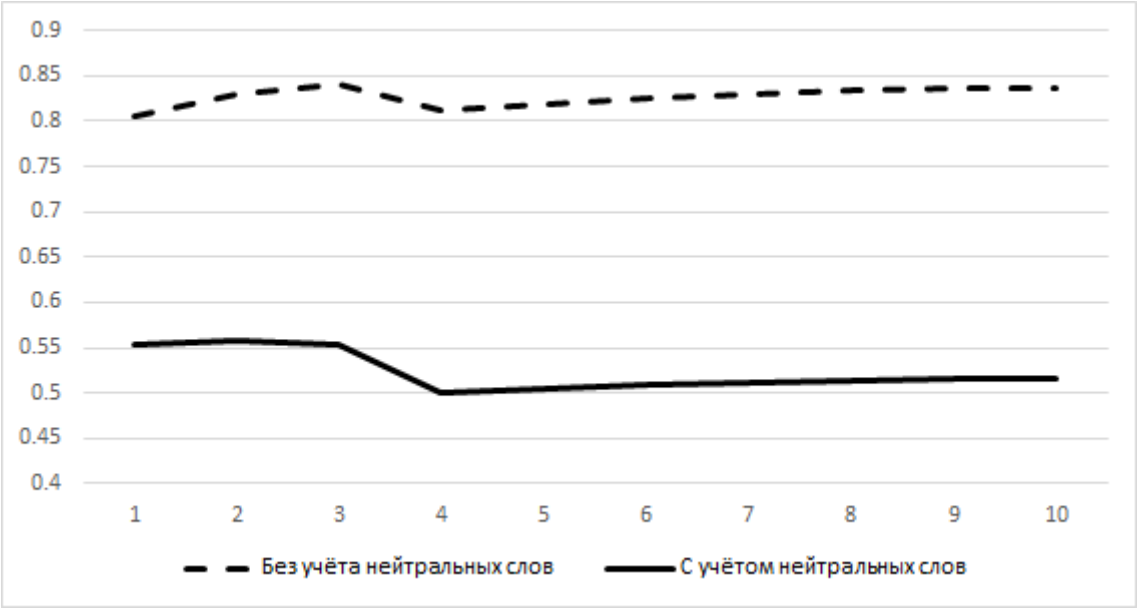

Рис. 11: Зависимость F_1-меры от параметра K без удаления “не-”

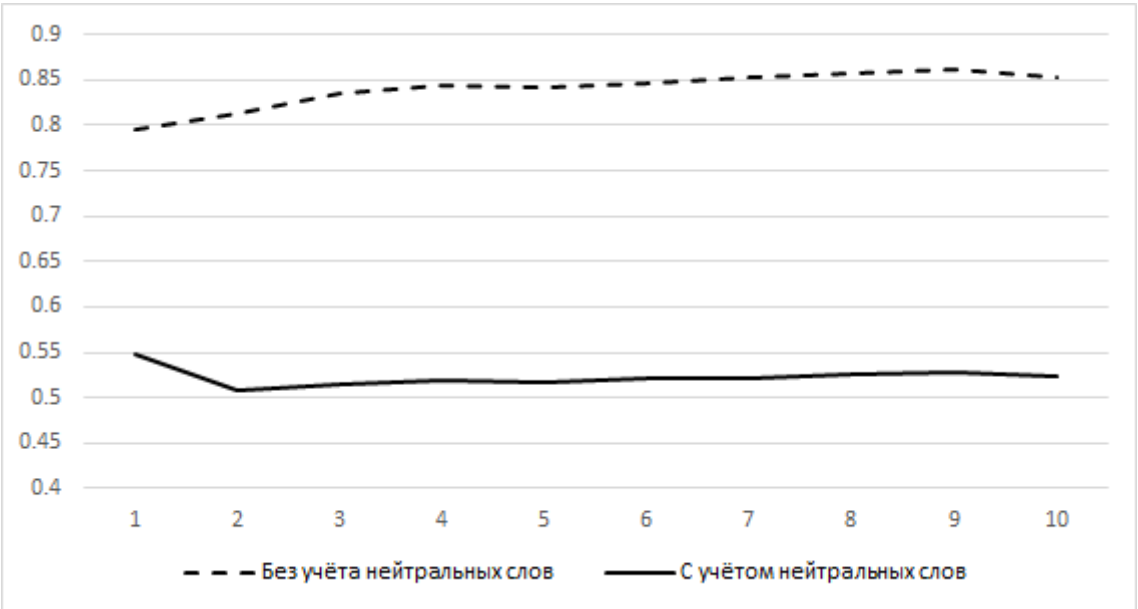

Рис. 12: Зависимость F_1-меры от параметра K после удаления “не-”

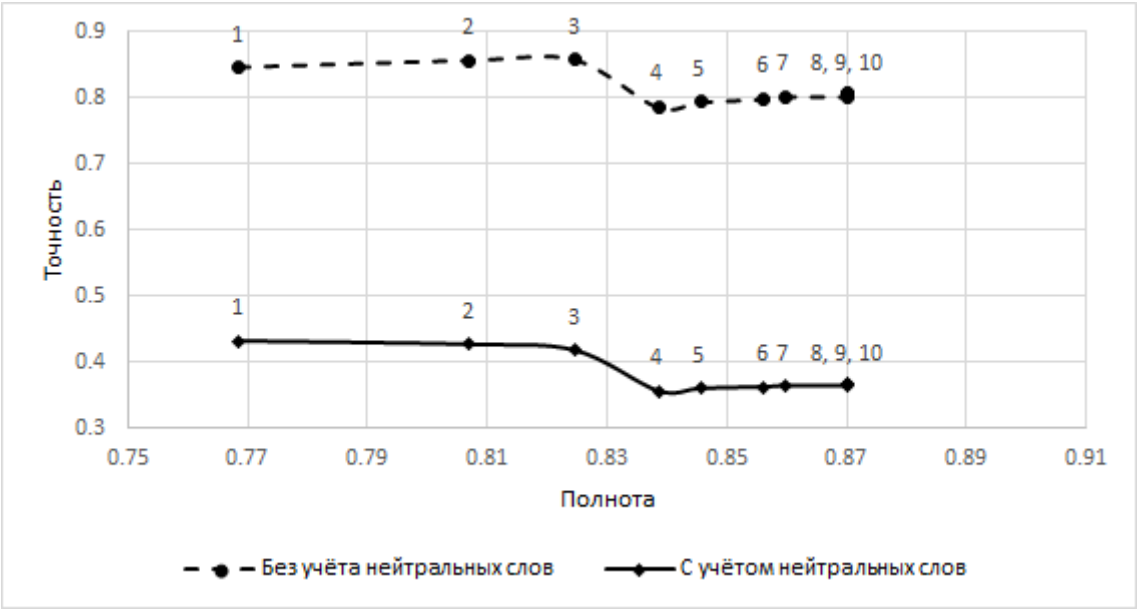

Рис. 13: Значения точности/полноты при различных значениях параметра K без удаления “не-”

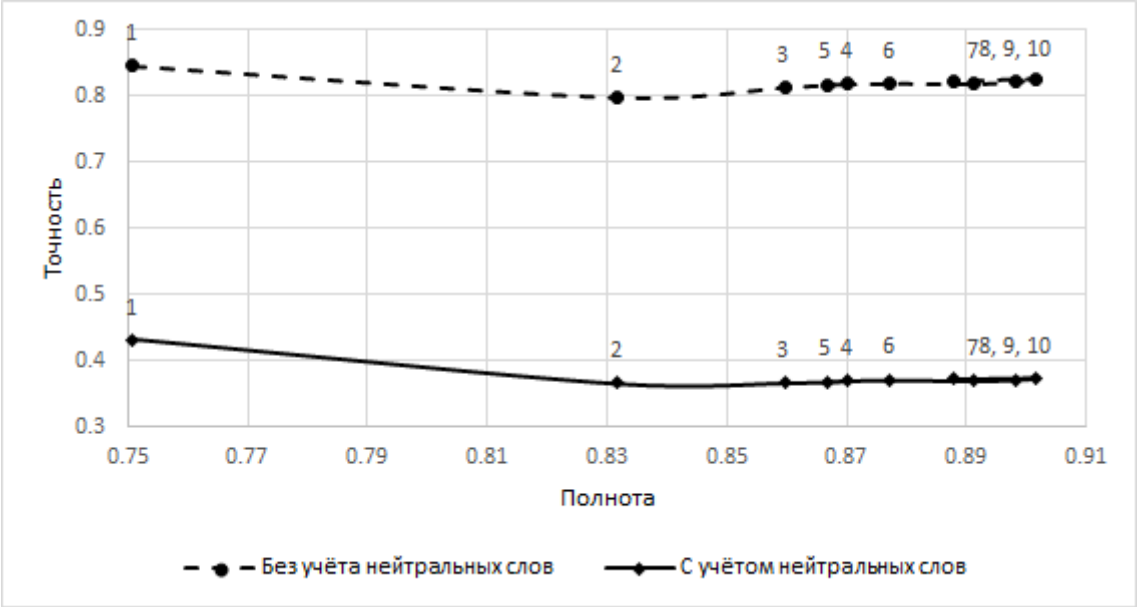

Рис. 14: Значения точности/полноты при различных значениях параметра K после удаления “не-”

6. Заключение

В данной работе предложен метод построения словарей позитивных и негативных прилагательных на основе анализа неразмеченных отзывов из произвольной предметной области. Для этого рассматривается и анализируется граф, построенный на прилагательных, встречающихся в тексте, как вершинах и синтаксических связях между ними в качестве рёбер. Для разделения прилагательных на положительные и отрицательные этот граф разбивается на два кластера с использованием в качестве начального множества "универсальных" прилагательных, эмоциональная окраска которых не зависит ни от предметной области, ни от контекста. В работе приводится сравнение нескольких реализаций алгоритмов как построения, так и анализа графа. Описанные алгоритмы обеспечивают построение словарей с точностью до 91% и полнотой 94%. Результаты работы опубликованы в [4]

Также проведен анализ существующих на данный момент и находящихся в открытом доступе русскоязычных тонально аннотированных словарей. Проведено сравнение имеющихся решений с полученными в данной работе результатами, а также оценка возможности и целесообразности применения каждого из этих словарей для анализа данных из другой предметной области для демонстрации существенных лексических и тональных различий между текстами различной тематики.

В качестве данных для экспериментальной оценки предложенных алгоритмов были выбраны русскоязычные отзывы об отелях. Однако описанный метод работает с неразмеченными текстами, а значит сформировать необходимую для работы алгоритма коллекцию текстов можно автоматически (например, с помощью краулера), без привлечения экспертов или оценщиков. Это позволяет использовать представленный алгоритм для произвольной предметной области. Кроме того, данный подход может быть применён к отзывам на других языках, поскольку для его реализации необходимы только морфологический анализатор, список сочинительных и противительных союзов выбранного языка, а также небольшое начальное множество "универсальных" с точки зрения

тональности слов, как то “хороший”, “плохой”, “ужасный”, “прекрасный” и т.д.

Список литературы

[1] Chetviorkin Ilia, Loukachevitch Natalia V. Extraction of Russian Sentiment Lexicon for Product Meta-Domain. // COLING / Citeseer. — 2012. — P. 593–610.

[2] Clematide Simon, Klenner Manfred. Evaluation and extension of a polarity lexicon for German // Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA) / Ed. by A Montoyo, P Martínez-Barco, A Balahur, E Boldrini. — Lisbon, Portugal, 2010. — August. — P. 7–13. — URL: http://dx.doi.org/10.5167/uzh-45506.

[3] Creating sentiment dictionaries via triangulation / Josef Steinberger, Mohamed Ebrahim, Maud Ehrmann et al. // Decision Support Systems.

[4] Dubatovka A., Kurochkin Yu., Mikhailova E. Automatic Generation of the Domain-Specific Sentiment Russian Dictionaries // International Conference on Computational Linguistics and Intellectual Technologies Dialog-2016. — 2016. — P. 146–158.

[5] Esuli Andrea, Sebastiani Fabrizio. PageRanking WordNet Synsets: An Application to Opinion Mining // ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic / Ed. by John A. Carroll, Antal van den Bosch, Annie Zaenen. — The Association for Computational Linguistics, 2007. — URL: http://aclweb.org/anthology-new/P/P07/P07-1054.pdf.

[6] Esuli Andrea, Sebastiani Fabrizio. Random-walk models of term semantics: An application to opinion-related properties // Proceedings of LTC 2007. — 2007. — P. 221–225.

[7] Harabagiu Sanda M., Miller George A., Moldovan Dan I. WordNet 2 – A Morphologically and Semantically Enhanced Resource // Proc.

SIGLEX 1999. — 1999. — URL: http://xwn.hlt.utdallas.edu/papers.html.

[8] Hatzivassiloglou Vasileios, McKeown Kathleen R. Predicting the semantic orientation of adjectives // Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics / Association for Computational Linguistics. — Madrid, Spain, 1997. — P. 174–181.

[9] Koltsova O. Yu., Alexeeva S. V., Kolcov S. N. An Opinion Word Lexicon and a Training Dataset for Russian Sentiment Analysis of Social Media // International Conference on Computational Linguistics and Intellectual Technologies Dialog-2016. — 2016. — P. 277–287.

[10] Loukachevitch Natalia V., Chetviorkin Ilia. Refinement of Russian Sentiment Lexicons Using RuThes Thesaurus // RCDL / Ed. by Lidia Kalmykova, Mikhail R. Kogalovsky. — Vol. 1297 of CEUR Workshop Proceedings. — CEUR-WS.org, 2014. — P. 61–65.

[11] Pang Bo, Lee Lillian. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts // Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics. — ACL '04. — Stroudsburg, PA, USA : Association for Computational Linguistics, 2004. — URL: http://dx.doi.org/10.3115/1218955.1218990.

[12] Segalovich Ilya. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. // MLMTA / Ed. by Hamid R. Arabnia, Elena B. Kozerenko. — CSREA Press, 2003. — P. 273–280.

[13] Wu Qiong, Tan Songbo, Cheng Xueqi. Graph Ranking for Sentiment Transfer. // ACL/IJCNLP (Short Papers). — The Association for Computer Linguistics, 2009. — P. 317–320.